Text S1: Sequencing, Alignment and SNP Calls

We used a mixed 454 and Illumina sequencing strategy. Genomic DNA from each House Finch strain was first nebulized and blunt-end ligated with Roche adaptors. Emulsion PCR was performed using the bulk library and these products were pyrosequenced on Roche 454 Gene Sequencer using FLX chemistry, using physical separation of isolates. Genomic DNA from the four newly sequenced poulty strains (Table S1) was sequenced by the Illumina method at the University of Utah Huntsman Cancer Institute.

Sequence data generated from the Illumina sequencing platform was used to make a multiple sequence alignment according to the following protocol. Raw sequences were first trimmed to retain only high quality sequence data. Trimmed reads over 25 bp in length were then aligned to the reference MG genome (AE015450.2), ignoring any ambiguously mapped reads using the CLC Genomics Workbench version 3.7.1 . Finally, bases were called from the consensus sequence in this alignment for each unmasked (see below) regions of the genome. Any basepair in these regions that differed from the consensus region was only called as a SNP if there were at least 4 reads at the position, if the base passed NQS (30/25) standards, and if at least 95% of all reads aligning at that position had the SNP base. To avoid errors due to runs of single bases, we also required that there be no more than 2 gaps or mismatches within an 11 bp window around the putative SNP in any read that was counted towards calling the SNP. This final criterion is effective at removing erroneous SNPs that are simply artifacts of poor sequence alignment. However, due to the high number of SNPs present in all four of our sequenced strains (10,007-10,729 differences in an alignment of ~756 kb), we would expect this to also exclude some legitimate SNPs that fell in windows with 2 or more neighboring SNPs simply by chance. For this reason, we generated a second dataset that varied this criterion by allowing up to 4 gaps or mismatches with in an 11 bp window around any position that differed from the reference genome. This dataset contained only 1% more SNPs than the original dataset, and we found that the conclusions in this paper were qualitatively unaffected by using this alternate dataset.

The sequencing data from the House Finch MG isolates generated on the Roche 454 platform data was aligned to the reference *Mycoplasma gallisepticum* genome [1] using the Mosaik aligner [2]. From this alignment a basepair for a strain was reported for each position in the reference genome if the following conditions were met. The majority base at that position had to have a center QC score of 30 or higher, and the 5 flanking bases on either side of it to each have a score of 25 or higher. We also required at least two reads before a base was called. A whole genome alignment was then generated by aligning the basepairs present from each sample at the same position in the reference genome. This whole genome alignment was then divided into sections and the alignments were manually curated by the authors. If during this process a region of the genome was found to clearly violate the Markov models we assumed for nucleotide evolution (e.g. a transposon insertion) or if the aligner was grossly in error (e.g. a section with a deletion or a slightly varying simple sequence repeat that confused the aligner) we either manually edited the relevant section or if it was not obvious what the correct alignment should be we excluded it from later analysis. We additionally validated our alignments by ensuring that the base we reported at every position in our alignments matched the base independently reported by a separate alignment algorithm. The other aligner we used was the run454Mapper program (part of the Genome Sequence FLX Data Analysis Software package available from Roche).

For both the Illumina and 454 sequencing data, we aggressively excluded repetitive segments of the genome that we believed to be inappropriate for SNP calls. In particular, the *Mycoplasma gallisepticum* genome contains a number of proteins that have a high degree of similarity with other proteins, such as the VlhA family of lipoproteins that constitute 10.4% of the reference genome, and also the Apr-E like proteins, transposases, CRISPRs, etc. This repetitive DNA is likely to undergo recombination and in some cases it is not possible to correctly align or assemble. To avoid artifacts introduced by these regions, we excluded any region of the genome over 100 bp in size that had over 85% similarity to another location in the reference genome as determined by megablast . In total, we excluded 228,875 bases (~23% of the reference, henceforth the masked segments) from our multi-isolate alignment and from our SNP calling protocol.

Alternate SNP Calling Protocols

To check the sensitivity of our results to our SNP calling method for the House Finch MG isolate data, we generated three alternate genomic alignments using different protocols and quality threshold levels. These datasets were all generated for only the unmasked portions of our genome, and are described below. All analyzes described in this paper were also performed with these alternate datasets and equivalent results were obtained.

a) Stringent Threshold - Broad Institute 454Swap SNP Calling Software

This dataset calls SNPs using software developed at the Broad Institute for 454 data [3]. The pipeline that uses this SNP calling software is completely independent of the other base calling methods we used. The quality threshold requirements for this program include:

a) A basepair needs to have reads that align to it coming from both the right and the left side.
b) A basepair must be represented by at least two reads that pass NQS thresholds.
c) No more than 33% of the reads at a position can disagree with the consensus base.

b) Moderately Stringent Threshold – Our working data set

This is our working dataset and was created as described in the preceding section. Using the protocol in which 2 gaps or mismatches within an 11 bp window were removed produced an alignment that was 756,552 bp in length. In addition, we also produced an alignment allowing up to 4 gaps or mismatches within an 11 bp window (with the level of diversity present in our poultry sample this many mutations could be expected to occasionally occur and so this may not always indicate sequencing errors). This alternate alignment of 756,574 bp contained only 1% more SNPs than the original method and use of this alignment did not affect any of the conclusions in the paper. This protocol also yielded an alignment of 738,209 bp when considering only the House Finch MG strains.

c) Moderate threshold -A variant of the moderately stringent data set

This dataset was produced exactly as above except there was no requirement that the data generated matched the data generated by the Roche 454 aligner.

Further SNP Dataset Validation

In addition to the checks described above, we also performed 76 traditional Sanger sequencing reactions of SNPs called in the House Finch MG dataset (Table S3) and confirmed that the SNPs were called correctly. Additionally, as our dataset is a composite of 454 and very high coverage Illumina sequencing data, we cross-validated our results by comparing these two datasets to each other and found excellent agreement between them (Figure S2, Table S4).

Text S2: Inference of mutation rate, recombination, times to common ancestry and population dynamics

Using BEAST v1.52 [4], estimates of times of common ancestors were obtained for both the 13-taxon alignment of 738 kb containing our 12 House Finch Strains and the reference genome (large alignment) as well as for 73-taxon LS-MSA of 1.3 kb, which included MG sequence data obtained from strains sampled between 1955 and 2000 [5]. To aid in the selection of the inference model and to ensure that the results based on the large alignment were qualitatively insensitive to inference model choice, we compared the estimates of the mutation rate obtained from a variety of different possible analyses. Since a population expansion was observed to occur over the sampling period, in all inference models considered we assumed a changing population size using the exponential skyline model [6], and also always assumed some form of the HKY nucleotide substitution model. Given these model choices, we also tested the effect of four additional choices, or factors, on our inference. One of these factors was the modeling choice for site heterogeneity, which we tried at three levels (HKY, HKY+Γ or HKY+Γ+I). We also varied the data by including and excluding the reference genome because it was sampled at a much earlier time point then the other strains and thus could exert a high amount of leverage on the rate estimate. Another factor was the multiple sequence alignment used, and we tested all three of our SNP calling datasets (Stringent, Moderate and Moderately Stringent). Finally, since the amount of sequencing data present for each of our strains varied, we tested whether the strains with greater coverage were biasing the results by running the analysis while allowing BEAST to average over partially observed sites, or by only analyzing sites with data for all strains.

In total this resulted in 36 ($3 \cdot 2 \cdot 3 \cdot 2 = 36$) different methods to infer the rate of evolution, and inference about the posterior distribution of the rate parameter was obtained for each of these methods from 10,000,000 MCMC samples. From this analysis, 2 of the 36 MCMC runs were unable to converge. These runs were performed with settings that essentially deprived the inference method of enough data to jointly infer the parameters in the model (e.g. the stringent dataset and the requirement that all strains have data present), and as a result the estimates were wildly varying and inaccurate (e.g. Median clock rates of 1.53e307) and the MCMC chains clearly failed to converge. A plot of the rate estimates from the 34 runs that could produce sensible results is shown in Fig. S4.

We concluded from this analysis that the rate estimate was robust to these model choices. However, the results reported in this paper are based on the model we believed to be the best, which included the reference strain to allow inference about its divergence time from the HF ancestor and, used the HKY+I model of substitution (when inferred, the posterior distribution of the Gamma parameter was identical to the prior because the low amount of diversity in the house finch MG meant there were not enough multiple mutations to estimate this parameter), and used our Moderately-Stringent dataset. For this model, we ran 8 additional chains starting from different initial trees and parameter settings, and checked that all converged to the same distribution. The results for this analysis gave an estimated mean clock rate of 1.02e-5 per year (95% HPD 7.95e-6 to 1.23e-5), an estimated date for the MRCA of the HF strains as having lived 19.2 (95% HPD 16.9 to 21.7) years prior to 2007 and estimates the common ancestor of the HF strains and the chicken reference to have occurred 599.2 (95% HPD 477.5 to 737.0) years

prior to 2007. We also used this analysis to estimate a skyline plot for the House Finch MG [6] (Figure 2).

In order to compare our rate estimates with the 73 taxon, 1.3kb alignment, we also estimated these quantities using BEAST, again from 8 different initial values, assuming a model of population change and using the HKY+G+I substitution model. This estimated the mean rate as 3.23e-5 (95% HPD 6.37e-6 to 6.239e-5), and the common ancestor of the HF strains and HF strains to have lived 456.7 (95% HPD 130.8 to 969.4) years prior to 2007. We caution that the estimates of the divergence dates from HF to poultry strains are very coarse and should be interpreted with caution, as the modern poultry industry likely alters the population dynamics of MG transmission in ways that may strongly violate the coalescent model assumed in BEAST.

Text S3: Evaluating the effect of frameshift and nonsense mutations

We looked for frameshift and nonsense mutations, which we refer to collectively as disrupter mutations. To find these mutations, we *de novo* assembled the 454 reads from our House Finch MG samples, and searched for proteins in the assemblies that had such mutations in them. As the 454 *de novo* assembler improves with increasing read coverage, we restricted our analysis to two of our samples with high sequencing coverage, AL_2007_37 and VA_1994. These strains also bookend the time period of this study. Because the TK_2001 strain is so genetically similar to the MG strains in this study isolated from the House Finches, we also searched assemblies generated from its sequencing data, using the CLC genomics workbench v.3.7.1. For all of these strains, we searched for disrupter mutations present in any of the genes annotated in the reference genome, except for genes with strong similarity to other parts of the genome as these genes are most likely to be misidentified or misassembled. We excluded any gene that was annotated as a VlhA gene or a transposon, or that contained a sequence over 100 bp in length that aligned to another area of the genome with over 85% identity as determined by megablast. By this method 105 of 763 genes (13.7%) were excluded.

For each gene of the remaining 658 genes we used the *de novo* reconstructed gene sequences to check for the presence of disruptor mutations. We considered a gene successfully reconstructed if we were able to find a matching segment amongst the assembled contigs that covered the entire gene (as determined by evaluating local alignments determined by Megablast), and that did not differ by more than 200 bp in size. We were able to find matches for all but 48 of the 658 genes (~93% recovery) in our VA_1994 strain, all but 41 in AL_2007_37 (~94%) and all but 20 (97%) in the TK_2001 strain. 17 of the genes were not recovered in VA_1994 and TK_2001 because they had been deleted along the branch leading from the reference MG strain to our isolates, while such deletions caused 29 genes to be unrecoverable in AL_2007_37. The remaining genes were excluded either because they were not completely covered by a single assembled contig, or in one case because an IS element was inserted into it.

To detect pseudogenizing mutations, each of the 617(AL) , 610 (VA) and 622 (TK) successfully reconstructed genes was translated to detect nonsense or frameshift mutations. This identified 85 possible mutations affecting 76 genes in AL_2007_37 and 99 possible mutations affecting 91 genes in VA_1994. For each of these mutations, we then examined the reads supporting them by evaluating the alignment of the reads to the reference genome in the .ace file produced by both the Newbler and Mosaik aligners. We found that many of the indel mutations were near homopolymers where the underlying reads often both supported and contradicted the presence of the relevant indel mutation. We disregarded all such ambiguous cases unless the reads supporting the presence of the indel outnumbered those contradicting it by 10. This criterion excluded 55 mutations in VA_1994 and 41 mutations in AL_2007_37. This left 44 mutations affecting 42 genes in AL_2007_37 and 44 affecting 43 genes in VA_1994. All of these mutations were shared between VA_1994 and AL_2007_37, except for two. One putative nonsense mutation along the branch leading to AL_2007_37 (reference position: 30,546) was found to have occurred in a gene that had already suffered a frameshift in the common ancestor

of VA_1994 and AL_2007_37. A second mutation was present in VA_1994, but because this area of the genome had been deleted in AL_2007_37, it could not be recovered from this sample.

Of the 45 disruptor mutations found, we excluded an additional 18 mutations because the mutation either occurred in a gene that had been annotated as a pseudogene, or because the mutation was actually supposed to be the wild type state of the gene. The later are likely due to sequencing errors or mutations in the reference genome and we determined this to be the case if the effect of the mutation was to merge two pseudogenes back into a functional protein, and if the mutation was present in all of our sequenced poultry strains as well. This left a total of 27 total disruptor mutations which we grouped into the following two categories. All mutations present in the VA_1994 strain were also present in the closely related TK_2001 strain, and some were present in the other poultry strains.

a) Extension Mutations

4 frameshift mutations had the effect of simply extending the length of the protein shown below. These mutations all occurred within the last 1% along the length of the protein, and although these changes do alter the amino acids towards the end of the protein, it is likely that these proteins remain functional.

Genes with extension mutations

Protein ID	Mutation Location	Mutation	Length of extension (aa)	Present in All Strains?
MGA_0809	132,809	A deleted	5	Yes
MGA_0812	135,135	T->A interrupts stop codon	5	Yes
MGA_1153	416,101	Single T deletion in TK_2001, AL and VA have a deletion of 2 "T"'s at this location	2	Only House Finch MG strains and TK_2001
MGA_0232	718,459	A deleted	11	All but TN_1996

b) Pseudogenes Formers

Excluding the extension mutations and mutations that disrupted the reading frame in one gene but merged it with an upstream coding sequence, we observed 23 mutations affecting 17 genes. These were distributed as 10 insertions, 10 deletions and 3 mutations of an amino acid coding codon to a stop codon. The mutations were often clustered in the same gene. There are 4 genes each of which had 2 mutations which would have disrupted the original reading frame, as well as 1 gene with 3 disruptor mutations. The remaining 12 genes were only disrupted by one mutation. The genes affected by these mutations are given in table S9.

Text S4: Transposon (IS) Movements

To identify areas of transposon insertion, and to determine if our isolates contained transposable elements in the same location as the reference genome, we developed a method to identify and annotate transposable elements from the 454 reads. Briefly, the method uses a querying strategy similar to BLAST to search for reads that contain sequences identified with the edges of IS elements. The method then annotates the portion of the read that belongs to the IS element, and maps the remaining portion of the read back to the reference genome in order to identify the location where the IS abuts a portion of the genome. The source code for the method is available from the authors upon request.

The reference genome contains members of 2 groups of IS elements. The first group present is identified by the ISFinder database [16] as belonging to the IS1634 family, and is represented in the reference genome by two complete transposases and one shorter fragment with high similarity to a complete IS element (we refer to such fragments as a scar). The second group includes members of the larger IS256 family, and is represented by ten transposases in the genome (although one of these is broken apart by another copy of an IS which has been inserted into it).

The transposases belonging to the first group (IS1634) have also been found in the genomes of other *Mycoplasma* species including *bovis*, *mycoides*, *hyopneumoniae* and *synoviae* [16]. Although this transposase seems to effectively persist in these other *Mycoplasma* genomes, it appears that no functional copy of this transposase remains in this study's House Finch MG strains. Of the two transposases annotated in the reference genome, only one was functional as the other had a frameshift mutation in it. Based on the Newbler assemblies of our sequence data, this particular transposase is even more degraded amongst the strains we sequenced. The first stop codon now appears only 30 amino acids into the gene in all the strains where we could confidently reconstruct it. The only remaining member of the family present in the reference genome, with the only functional transposase, is gone entirely from the House Finch samples we sequenced. It appears that this remaining functional transposase recombined with one of its scars, leading to a large deletion and the destruction of this last functional copy.

In contrast, the second group of transposases, belonging to the IS256 family, has been active during the divergence from the most recent ancestor of our samples and the reference genome. In the reference genome, this group is represented by 10 transposases and 3 small scars. However, in our samples, only 4 of these 10 IS elements are present. Three IS elements in the reference genome had not been inserted by the time the reference strain and the strains in our samples diverged, and three of the other IS elements were located in a region of the genome that had been deleted in the lineage leading to the common ancestor of all of our samples.

Along the branch leading to the common ancestor of all our samples, this element inserted itself into 6 new locations (Table S8). Each of these insertions shown was present in every one of our HF samples, and no sample had any insertion that was not present in the others. Of the 6 IS element insertions, 4 were in intergenic regions, which given the density of genes in the reference genome is highly unlikely ($p < 0.003$). A likely explanation for this bias is that selection is filtering out insertions that destroy functioning genes.

Text S5: Searching for Novel Genes in the House Finch MG isolates

This study relied on comparing the assembled genomes of the 12 House Finch MG isolates with that of an annotated reference strain. Given the amount of divergence between the reference and our samples, it was important to determine if using this reference genome would prevent us from analyzing additional gene sequences that were not present in the reference genome but that could provide additional information for this study. To investigate the presence of potentially novel genes in our House Finch isolates, we searched the contigs generated via *de novo* assembly for DNA sequences that could not be mapped back to the reference genome. To do this, we megablasted all of the assembled contigs against the reference genome, and examined any section of a contig sequence longer than 100 bp that could not be mapped to the reference genome. To maximize our chances of detecting any novel sequences in the assembled contigs, we examined the contigs generated from our high-coverage VA_1994 and AL_2007_37 strains. We also pooled all of our 2007 samples for *de novo* assembly and investigated the contigs that were generated from this meta-sample.

Few if any novel DNA sequences were found and arguably none were truly unique because they all had strong similarities to members of either the VlhA or AprE-like proteins present in the reference genome. Of the sequences that failed to align with high similarity to the reference genome, several sequence segments ranging in size from 100bp to 2.1 kb could be identified as similar to a VlhA region by BLAST or BLASTX. However, the largest segment of these that aligned with less than 80% similarity to a portion of the reference genome was only 1.6 kb in size, and a translation of this sequence revealed that it contained a Vlh-A type gene. Similarly, there was a ~500 bp segment that could not be mapped to the reference genome, but this segment was flanked by ~3.3 kb of DNA sequence that had between 66-70% similarity with the other AprE-like proteins present in the genome. These results were consistent for all of the assemblies tested. Given the difficulty in reconstructing these repeat-rich loci and their unsuitability for calling SNPs, we did not pursue these segments further.

Text S6: Detecting recombination

Despite the small amount of genetic variation segregating amongst our House Finch *Mycoplasma* samples (only 412 SNPs), it is not possible to build a single phylogenetic tree with no homoplasies from this data. Similarly, although our poultry strains contained many more SNPs between them, one still cannot infer a single phylogenetic tree that has much more support than alternate trees. The reason for this is not that the SNPs provide very little information about phylogenetic relationships, but rather that many SNPs provide information that is in conflict with the information provided by many other SNPs as determined by the four gamete test. This type of behavior is expected if genes are flowing horizontally as well as vertically in a population, and so we formally tested for the presence of recombination in our dataset.

A plethora of tests are available to detect recombination in sequence data (see ref [7-9]). However, because many of these tests examine the same fundamental signal of recombination, such as the physical clustering of phylogenetically concordant SNPs, they commonly yield qualitatively similar results when performed on the same dataset. To detect recombination in our combined House Finch and poultry strain dataset, we used the pairwise homoplasy index test[7] as implement in splitstree4 [10]. Examining the entire data set, this test found a statistically significant signal of recombination ($p < 1e\text{-}9$). This signal comes predominantly from the four newly sequenced poultry strains because there is not enough genetic variation to make the test significant when only the house finch strains are considered. However if we apply to the house finch MG strains the homoplasy test by Maynard-Smith and Smith [11], which is found to perform well in situations of low nucleotide diversity [9], we still obtain a significant signal for recombination. This test differs from the pairwise homoplasy index test, in that rather than looking for spatial clustering of phylogenetically concordant SNPs, it instead asks if the number of homoplasies observed on a tree is particularly large given the number of mutations on the tree and the number of sites that were available to mutate. To implement this test, using the dnapars program in the phylip package we first found a parsimonious tree for the House Finch isolates while including the reference genome as an outgroup, and then counted the number of homoplasies that appear only within the clade of House Finch MG isolates. This identified 13 homoplastic mutations out of a total of 412 variable sites in an alignment of approximately 756.5 kb of DNA. Intuitively, it seems extremely unlikely to see so many homoplastic mutations given the large number of sites that were available to mutate. However, exactly quantifying how unlikely this is complicated because different sites in the genome evolve at different rates, so that homoplasies are much more likely to appear at some sites than others. To get around this issue, the original paper describing the test proposed reducing the total number of sites in the alignment down to a smaller number of effective sites. This paper described a heuristic method to estimate how much one should reduce the alignment size, but this method required a sequence from a distant outgroup that fulfilled a difficult set of assumptions. In practice, since having such an outgroup and trusting that it satisfies the assumptions is rare, many researchers simply take the effective number of sites to be equal to 0.6 multiplied by the total number of sites in the alignment. This 0.6 value was selected as a conservative choice when it was first used in a paper comparing different methods of testing for recombination [9] because it was much less than the inferred values in the original paper (0.73-0.83) and because 0.6 was given as the lower limit for a believable estimate in that same paper during a discussion of different methods to estimate the effective number of sites. Since then, this 0.6 value has been used widely in other papers and is

the default setting implemented in software packages that implement tests for recombination such as START (Sequence Type Analysis and Recombinational Tests, [12]). Suffice is to say, it is clear that there is some ambiguity in how best to determine the effective number of sites that are available to mutate, particularly when there is not complete sequencing for every strain, and as a result the homoplasy test could be considered overly conservative or subjective depending on one's prior beliefs. However, because the observed number of homoplasies in our dataset is so unlikely, we can confirm that it is extremely unlikely even given a wildly conservative set of assumptions. To determine the effective number of sites we used in our test, we first dropped the total number of sites in the alignment from 756,552 bp, down to the total number of sites where all the strains had data present which was only 273, 482 bp. This is obviously an overly conservative reduction as the vast majority of sites in the alignment had data for a majority of strains. Next, instead of applying the standard 0.6 correction to this reduced number, we applied a much more stringent criterion of 0.2, leaving us with 0.2 x 273482 = 54,695 effective sites, or only 7% of the original alignment length. We then estimated the probability of observing 13 or more homoplasies by simulation. Of 1 million simulations, the highest observed number of homoplasies was only 9, and we thus estimated our p-value as $p<1e\text{-}6$. However, the probability of observing so many homoplasies is almost certainly lower than this bound, not only because every assumption we made is expected to increase the p-value, but also because in our dataset two sites needed to convergently mutate not twice, but three times each in order for them to be in agreement with the tree. Since the homoplasy test treats all homoplasy counts as equivalent, even though repeated homoplasies at the same site are particularly unlikely, this again introduces a conservative bias into the test. We also note that the homoplasy test does not consider that a mutation at a site need not produce the exact same basepair each time, as there are three basepairs available to mutate to, which introduces yet one more conservative bias into the test.

Having established that *Mycoplasma gallisepticum* is a bacterium that recombines, we next sought to characterize the nature of recombination in this organism. To bookend a continuum with a dichotomy, recombination between microorganisms can be described as either chunky or smooth. Recombination is chunky when the recombination rate is much lower than the mutation rate, so that the genome is filled with large blocks of easily identifiable DNA that have a shared history that is in strong disagreement with the phylogenetic pattern exhibited by other sections of the genome. In contrast, recombination is smooth when the recombination rate is nearly equal to or greater than the mutation rate, in which case clusters of phylogenetically concordant SNPs tend to be much smaller and correctly delineating a specific section of DNA that has not recombined since the last common ancestor is impossible to do with any reasonable certainty. We therefore looked at the size distribution of phylogenetically concordant chunks to examine the extent to which the statistically significant finding of recombination was due to a few large blocks, or many smaller blocks.

To do this, we systematically determined the size distribution of phylogenetically concordant genomic segments in our sequenced isolates by implementing a recursive method that assigned each possible basepair in the genome to a phylogenetically concordant segment. Our method, illustrated in Fig S3**,** proceeds as follows. First for the strains under study we enumerate all possible unrooted trees. Next, for each phylogenetically informative SNP in the genome, we determine which trees are compatible with and incompatible with the pattern of variation shown at that SNP. In the next step, for each tree we determine all blocks in the genome that are in

agreement with that tree by assigning regions of the genome with consecutive compatible SNPs to single continuous block, and allowing half of the genome between a concordant SNP and a disconcordant SNP to be included in the block. Finally, all trees are examined to determine which has the largest block, this block is assigned to the tree, then the segments in each tree are updated to account for this, and this is repeated until every position in the genome is assigned to a block.

To implement the recursive method on our dataset, we first disregarded the data from the House Finch MG isolates. The House Finch MG isolates have too little genetic variation to usefully determine spatial patterns of recombination and were nearly genetically identical to the TK_2001 poultry isolate. By only using the MG poultry isolates and the reference genome, it is possible to work with the full enumeration of possible unrooted trees as there are only 15 and so we could avoid approximate and heuristic methods. The distribution of sizes of phylogenetically concordant blocks is shown in Fig S4, which also displays a distribution obtained by randomly rearranging the patterns of genetic variation shown at each SNP to different positions in the genome.

Fig S4 shows that the signal of recombination in our dataset is not due to a few rare transfer events, but that these genomes are reasonably mixed, as there are a large number of sizable concordant blocks that are in agreement with different trees. We note that we can also test for recombination by creating permutated datasets that keep the position of SNPs fixed and randomly reassigning the patterns of genetic variation shown at each SNP. If spatial clustering is significant, then the number of blocks required to assign the entire genome to a segment should be much less than the number required in a permutated dataset. Fig S5 shows the distribution of blocks required in 2,600 random permutated datasets, and as expected the total number of blocks required is much greater than that required in the actual dataset, again indicating that recombination is statistically significant with a vanishingly small p-value.

Text S7: Effect of recombination on the estimated substitution rate and demonstration of true temporal signal

The presence of recombination could bias our estimate of the substitution rate as inferred from BEAST. The MCMC algorithm used within BEAST proposes and evaluates the parameters in an evolutionary model based on a single phylogenetic tree. However, in the presence of recombination, there is no single phylogenetic tree that represents the history of all of the genomes sequenced, and this discrepancy between the biological reality and the inference model could affect our results. Although we hope future computing developments that use the ancestral recombination graph approach will eventually solve this problem by allowing the current Bayesian inference approaches to account for recombination, at present there are no available methods to systematically perform simultaneous inference of the posterior distributions for all the evolutionary parameters in circular-genome datasets as large as ours. However, despite this difficulty, it is clear that the single best point estimate for the mutation rate will always be on the order of 10^{-5} per site per year, and that given a number of well supported assumptions, that the interval of uncertainty around this estimate will encapsulate this rate to within an order of magnitude.

To demonstrate that our conclusions are robust to the presence of recombination, we note that a simple method of inference which is less affected by recombination gives virtually identical results. A naïve estimate of the mutation rate can be obtained for any two pair of sequences simply by dividing the number of mutations that appear between the earlier sample and the later sample by the amount of time separating the two samples. This method does not require that no recombination has occurred, however it does require that every element in the present genome has diverged for an equal amount of time from the genome it is being compared to. For example, if two genomes are thought to be diverged by 20 years, but lateral gene transfer (LGT) has introduced into the genome some segments that are diverged by over 40 years, than these segments will bias the mutation rate upwards if the 20 year period is used for the entire genome comparison as these more diverged segments likely contain more mutations. Although this makes LGT events typically problematic for these simple rate estimates, due to the host-shift observed in this system, the assumption of equal divergence times for all segments of the genome that differ between our older and newer samples is likely met. Based on the genetic evidence in this paper, the host-shift appears to be a single founder event that created an isolated population with no additional inputs from the source poultry population. This implies that even if recombination is ongoing between the 1994 and 2007 samples, since all of the strains in the population had a recent common ancestor near 1994, any segments introduced by LGT between 1994 and 2007 should be as diverged as the segments they are replacing.

While we would not expect the value of this estimate to be biased by recombination, this naïve estimate is biased towards a higher rate because simply dividing by the difference between the dates when strains were sampled does not account for the time between the last common ancestor of the two samples and the time of initial sampling, which is additional time during which mutations could appear. However, the nature of our data is such that this bias is very small, and any realistic correction for this bias does not substantially change the inference. The reason for this is that the most common ancestor of all of the House Finch MG strains was almost certainly present near the time of our initial sampling period, as supported by three lines

of evidence. First the epizoodic was very well documented as beginning in 1994 by a wide variety of observers, and despite ample opportunity there were no reports of MG infection in House Finches before this date. Second, and in agreement with 1994 being the first year when MG infected House Finches, in a broad sampling of MG from a variety of host species, all of the House Finch MG strains were genetically identical, despite a large amount of diversity in the poultry population, indicating a recent founder event (Fig S1). Finally, our genome level sequencing of the 1994 strains provides additional evidence for this interpretation. The 1994-1995 samples are characterized almost exclusively by singletons (Table S2), indicating a recent common ancestor and population expansion, and therefore a small bias in the naïve estimate. Therefore, given that there was a bottleneck in the founding of the house finch MG strains, the excess time not accounted for by the difference in sampling times is expected to be very small, on the order of a few months compared to the 13 year interval between the 1994 and 2007 samples, meaning that this naïve method, equivalent to a Poisson regression, will provide a very good estimate of the substitution rate. Evaluating this naïve estimate over any given pairwise comparison of 1994 and 2007 strains we get an estimated rate of 1.35-2.36 X 10^{-5} with an average of 1.7 X 10^{-5}. Although calculating an interval of uncertainty around these estimates is dependent on assumptions about the evolutionary process, one assumption that is uninfluenced by the effects of recombination is to assume that mutations are introduced into the genome as a constant Poisson process. With this assumption, the lower interval for the 95% confidence interval of our mutation rate is still on the order of 10^{-5} for the strains in this study. Although violations of a constant Poisson process are some of the most frequent findings in the field of molecular evolution, correctly identifying and modeling such deviations would require much broader sampling of bacteria then this study, or any other published study we are aware of, could provide. However, all indications are that such violations are not large in magnitude (Fig 4), and even if the width of the 95% confidence interval for the rate estimate assuming a Poisson process is doubled in size, the lower bound of the confidence interval is still approximately 10^{-5}. Therefore, we find no plausible violations of the model large enough to substantially alter our rate estimate more than an order of magnitude.

Additionally, as a simple test and demonstration that our data do contain a true temporal signal and the estimated rate is also not an artifact of the BEAST analysis, we used the program Path-O-Gen to evaluate the clock like nature of the data. An ML tree without an assumed clock was first estimated using the program PhyML [13] and the HKY substitution model used in our BEAST analysis. The regression in Path-O-Gen obtained an estimated rate of 1.45 X 10^{-5} using the default root for the tree (R^2 = .68) and it estimated a rate of 9.6 X 10^{-6} (R^2 =.92) using the best-fitting root, confirming that our Poisson regression results and the BEAST analysis are in agreement with this separate method of estimation. Finally, we also performed a randomization test as described in [14] by randomly reassigning the dates of all of our House Finch strains and rerunning our BEAST analysis We performed this randomization 20 times and each time obtained an HPD interval for the rate that did not overlap with our current estimate and was below our current estimated interval.

Text S8: CRISPR Analysis

To annotate the CRISPR elements in our 454 data and Illumina data we designed the program that computationally reconstructs a CRISPR locus from the sequence data, and simultaneously provides visualization tools that allow the user to validate the computational reconstruction, detect polymorphism, and manually check any ambiguities that may appear during the reconstruction. Each read generated from a sample is checked to see if it contains a sequence similar to the CRISPR repeat present in the ancestral genome. Any reads that contain such a sequence are selected into a subset of reads for further inspection. For each read in this subset, a dynamic alignment algorithm is used to identify the portions of the read that belong to the CRISPR repeats, and by exclusion, those portions that belong to the CRISPR spacers. Spacers that are exactly the same, or that appear to only differ due to sequencing errors, are then identified as spacer families, and these families are each given a numeric name determined by their (essentially random) order of discovery. Finally, as in many modern genome assemblers, the complete CRISPR locus is reconstructed through means of a graph. Each spacer family represents a node that can be placed either upstream or downstream of other families that appear in the same read as themselves. The program constructs this graph, determines the order of the observed spacer families, and plots it in simple format for the user.

We used this method to reconstruct the complete CRISPR locus in all 16 of our samples (Table S12). Of the 61 unique CRISPR spacer regions present in the reference genome, none are present in our samples, which collectively have gained a net total of 47 unique spacers since the time of their divergence from the reference genome. Table S8 shows the number of CRISPR spacers in each strain. Our GA_1995 sample has a copy of every spacer found in every other HF MG strain as well as TK_2001, and thus all other House Finch MG genotypes can be represented by deleting or duplicating the CRISPR spacers found in this strain. As such, the CRISPR array in each strain can be represented as a vector of discrete character states. Because adjacent CRISPR spacers are likely to be lost by the same deletion events, we reconstructed the CRISPR tree (Fig 5) using a parsimony method that always allowed deletions of neighboring CRISPR regions to be scored as single events. Following this assumption, we grouped strains into clades based on the presence of shared deletions or duplications. In instances where two or more equally parsimonious explanations could be provided for a pattern of deletions, we represented this ambiguity in the tree (for example the pattern of deletions shared by 2001 and 2007 can be equally well explained by having or not having these groups share a common ancestor to the exclusion of other strains).

Like the SNP phylogeny, the CRISPR phylogeny is consistent with a single origin of the epizootic and implies periodic replacement of the standing genetic variation in successive cohorts (2001, 2007). Although the deletions and expansions of most of the CRISPR spacers shown are likely due to strand slippage or recombination between the CRISPR repeats, the loss of the CRISPR spacers at the start of the locus in the 2007 strains is part of a much larger deletion of 12.7 kb that is unique to these strains and involves an alternative mechanism (deletion 4 in Table S7).

A recent investigation of CRISPR spacer repeats in *Yersinia pestis* found that a majority of spacers found in the CRISPR array originated from other areas of the organisms genome,

indicating that the CRISPR loci might be involved in regulating intra-genome dynamics, such as controlling gene expression levels or IS/prophage proliferation[15]. To determine if this was also true for the spacers we observed in *Mycoplasma gallisepticum,* we blasted each of the 302 unique spacer sequences that we found against the reference genome (blastn, with parameters "*-W 7, -e 1, -F F -r 2"*), and looked for any sequences that had an alignment score over 40 (equivalent to a ~66% match) to the reference genome. This analysis showed that 3 of the 302 spacer sequences were perfect matches to other portions of the genome, meaning they likely originated from proto-spacers within the MG genome. One spacer found within the reference genome was derived from a sequence within a hypothetical membrane protein (annotated MGA_0908), and another from the reference genome was a perfect match to a segment of DNA topoisomerase IV subunit A (MGA_0056). The third perfect match was from the CK_1996 strain, which contained a spacer sequence derived from a VlhA.4.01 lipoprotein gene (MGAH_0966).

To determine if any of the other CRISPR spacers were similar to any previously sequenced organisms, we blasted each of the 302 spacers against the NCBI 'nr' blast database and examined the top hit after excluding any hits to MG genomes. The top scoring hit only had a score of 50, and the top 5 hits were to *Schistosoma mansonis,* Human*,* and Zebrafish, leading us to conclude that there we no significant matches. Based on these comparisons to known DNA within and outside of the MG genome, we concluded that the source of the CRISPR spacers in this study is predominantly was from previously unstudied organisms.

Fig S1: Broad sampling of House Finch and poultry MG strain diversity.

To understand the broad phylogenetic diversity of House Finch and poultry MG strains, guide our choice of poultry strains for genomic sequencing and compare mutation rates in the HF and poultry MG population, we used DNA sequence data from Ferguson et al. [5] to generate a multisequence alignment for 82 MG strains collected from four host species (Turkey, Chicken, House Finch and Gold Finch). This data, henceforth the **L**arge **S**ample **M**ultiple **S**equence **A**lignment, **LS-MSA)** was composed of four gene fragments (from pvpA, mgc2, gapA and an unnamed surface lipoprotein) that when concatenated yielded approximately 1.9 kb of sequence data per strain (with the exact length of each strain varying due to small indels). We added to this dataset sequences for 8 of the 12 House Finch MG strains sequenced in this study that had complete coverage for these gene fragments. The four strains from this study not incorporated into the dataset (TN_1996, GA_1995, AL_2001_53 and AL_2007_05) were excluded because there was not enough sequencing data to accurately assemble the relevant fragments. We also excluded 3 strains from the original work[5] where we could not identify the host-animal species, leaving 82 strains in the final multiple sequence alignment. In this alignment, all the House Finch haplotypes were identical, except for the 2007 strains that differed from the others at two adjacent nucleotide positions.

Certain sections of the gene fragments in the LS-MSA were polymorphic due to insertions/deletions of tandem repeats, and because there is no clear criteria by which to assign the locations of these repeats in an alignment for phylogenetic purposes, for analysis purposes we reduced the ~1.9kb of sequence down to1,363 bp that could be confidently aligned.

Fig. S1. Phylogenetic tree of 82 avian MG strains inferred from four concatenated gene-segments, totaling 1,363 bp, using Neighbor-joining in PHYLIP. Due to recombination in *Mycoplasma gallisepticum*, this single tree may not be completely representative of the organismal history of the strains from which the gene segments were sampled. However, the pattern showing poultry hosts interspersed amongst the leaves of the tree and high diversity within the MG population is also present in neighbor-joining trees separately inferred for each individual gene fragment, consistent with frequent host-shifts by MG. Strain K4366GF97_10 is from an American Goldfinch (*Carduelis tristis*), also a songbird and the chicken reference strain used to obtain the reference genome is R63_44.

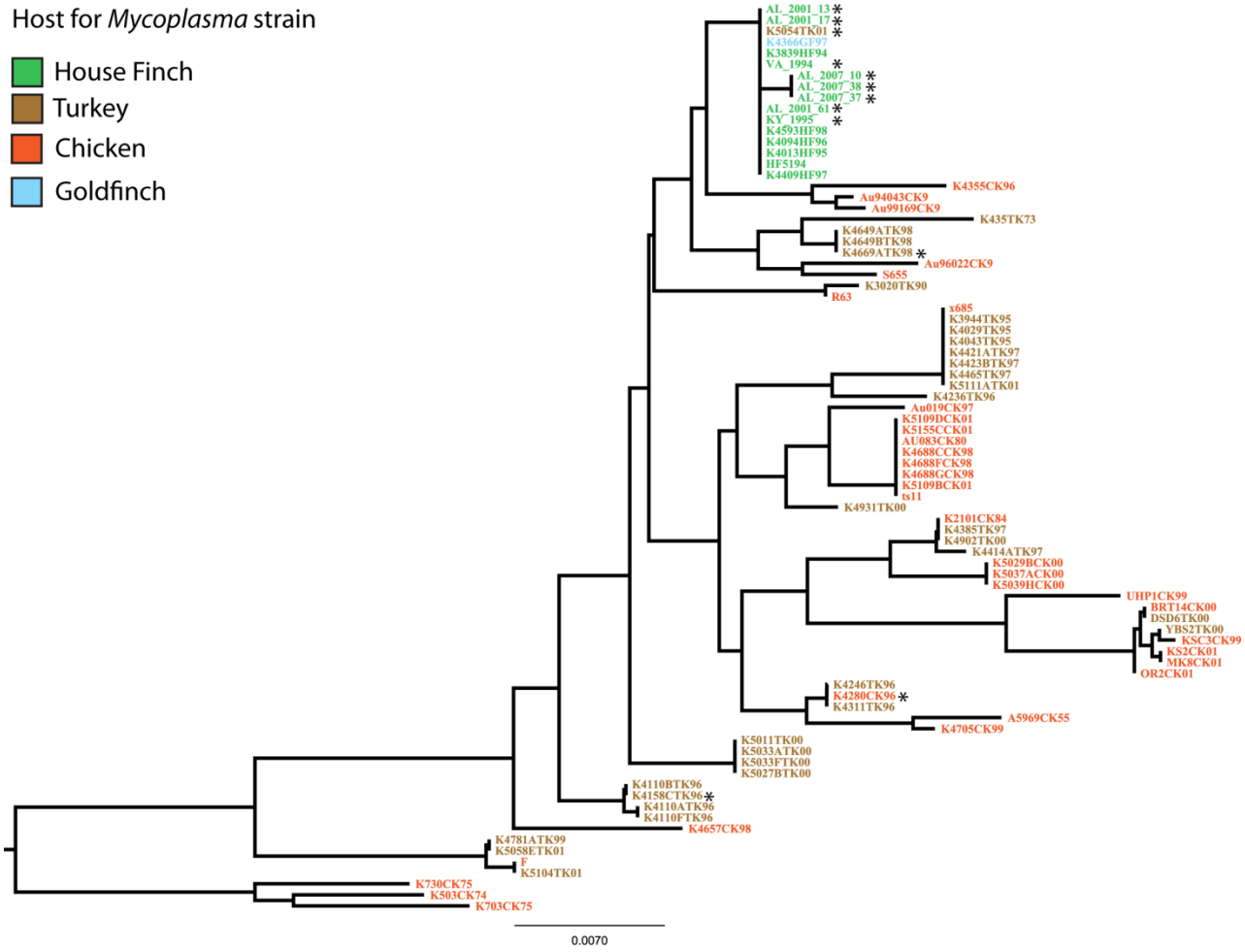

Fig S2: Cross Validation of the 454 Sequencing Data with the Illumina Sequencing Data

Our dataset provides an opportunity to validate the SNP calls made with our 4X-19X coverage 454 data for the House Finch MG isolates by using the SNP calls made with the 294X coverage Illumina data that was generated for TK_2001. TK_2001 and the House Finch MG isolates (particularly the pre-2001 isolates) are nearly genetically identical, and SNPs for both strains were called relative to the much more distantly related strain that was used to generate the reference genome. As outlined with the unrooted tree shown in Fig S1 this means that most of the SNPs called for each of the House Finch isolates should also be called for the TK_2001 strain, with any unmatched SNPs likely due to either genetic divergence between the two strains or SNP calling errors.

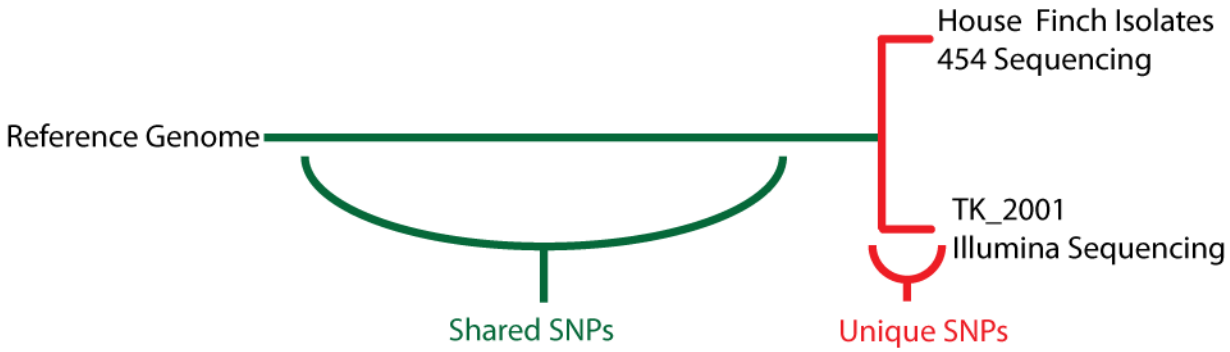

Fig. S2 Schematic representation of the cross validation approach.

The House Finch isolates that were sequenced with the lower coverage 454 data are very closely related to the TK_2001 strain sequenced with Illumina technology. Since SNPs in both TK_2001 and the House Finch isolates were called against the reference genome, comparing the high coverage TK_2001 Illumina strains to the House Finch Isolates allows us to estimate the total number of SNPs that could either be due to false positives or genetic divergence between the strains.

The results of this comparison are shown in table S4**.** For our most stringent threshold, of the up to 6,461 SNPs that were called in our pre-2001 House Finch isolates, 99.7% of the SNPs called with the 454 data were also called with the Illumina data. This bounds the false positive rate for SNP calls in the 454 stringent data at 0.3%. However, we believe that this unmatched 0.3% is due to true genetic divergence between the strains and not sequencing errors, as these SNPs are very well supported. For example, all 21 SNPs in VA_1994 that did not match TK_2001 were supported by at least 9 reads that contained the variant, and often many more.

Table S4 documents the robustness of our population genetic estimates on variations in SNP calling protocol, leading only to minor variations (~1%) in the false positive rate for our SNP datasets. This shows that almost all of the uncertainty in estimating the mutation rate from these genomes is due to the inherent sampling variability that naturally results from the stochastic process that generated them and is not due to any variability that comes from calling SNPs in

these genomes. Additionally the ratio of polymorphic to conserved sites is equivalent across all three datasets.

Fig. S3. Illustration of the recursive method used to assign segments of the genome to phylogenetically concordant blocks.

At the initialization of the algorithm the phylogenetically informative SNPs in the genome (x's in the diagram) are used to determine continuous segments that are in agreement with all possible trees. Sections of a genome in agreement with a particular tree are shown as solid colored lines over that genome segment. Note that any one SNP can be in agreement with multiple trees. If only one of two adjacent SNPs are in agreement with a tree, then half of the distance between the two SNPs is assigned to the concordant segment.

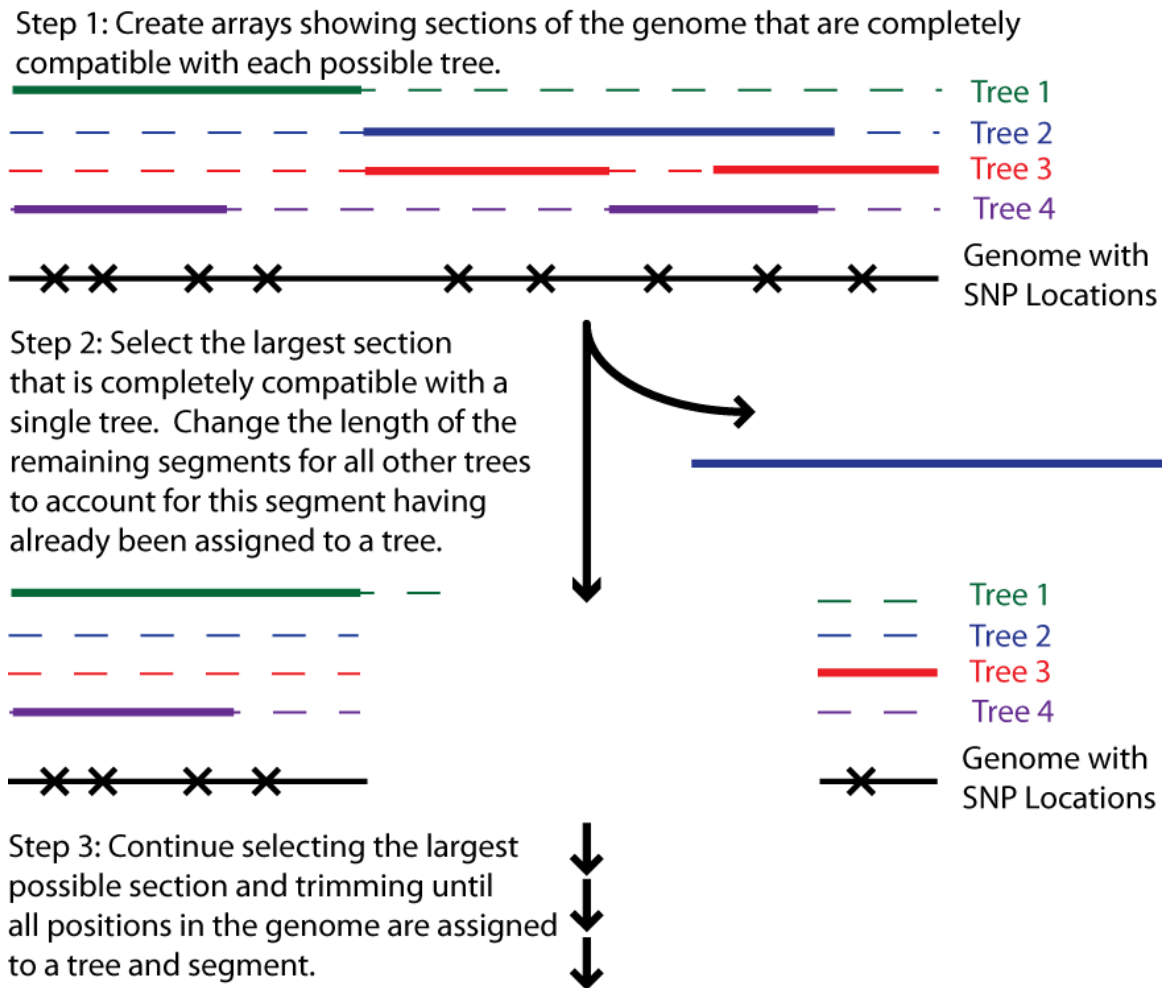

Fig. S4. Distribution of the number of phylogenetically concordant segments in the genome and in a dataset obtained by a single random permutation of the SNPs. Block sizes are in bp.

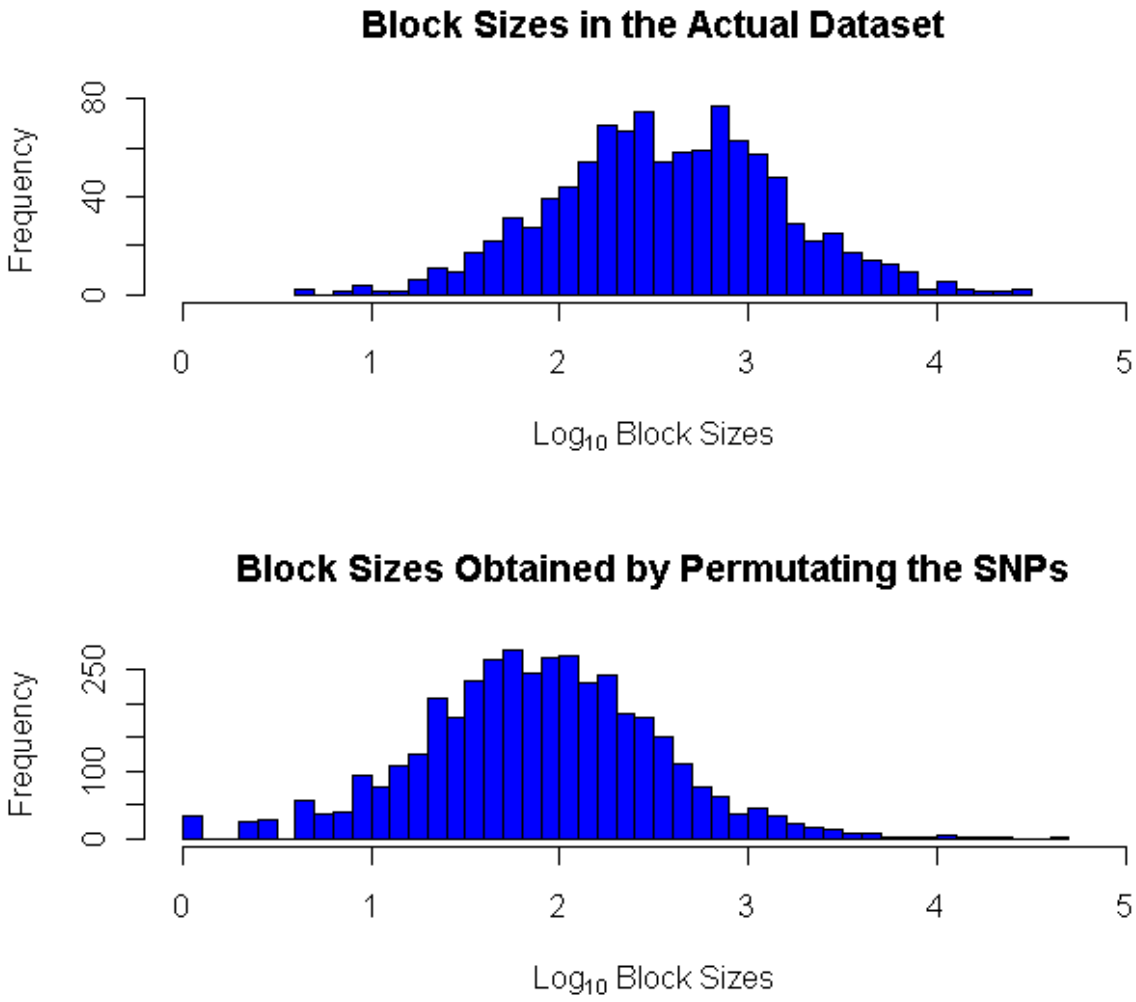

Fig. S5. Distribution of the size of phylogenetically concordant segments in the genome and in a dataset obtained by repeatedly creating permutations of the SNPs.

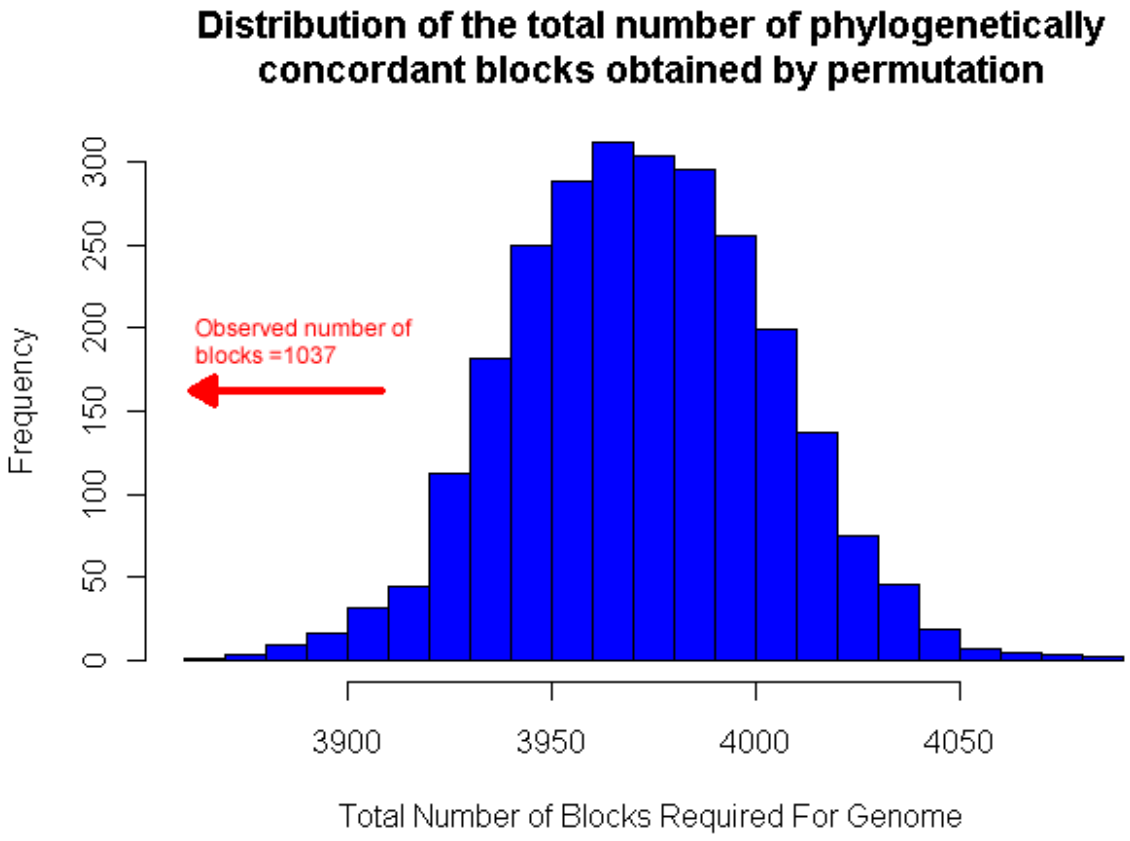

Fig. S6. 95 % HPD intervals of the rate estimated in BEAST using our actual dataset, as well as 20 permutations of the data where the dates on the tips are randomly reassigned. The interval for the true dataset is shown in red, and the randomized datasets are shown in blue.

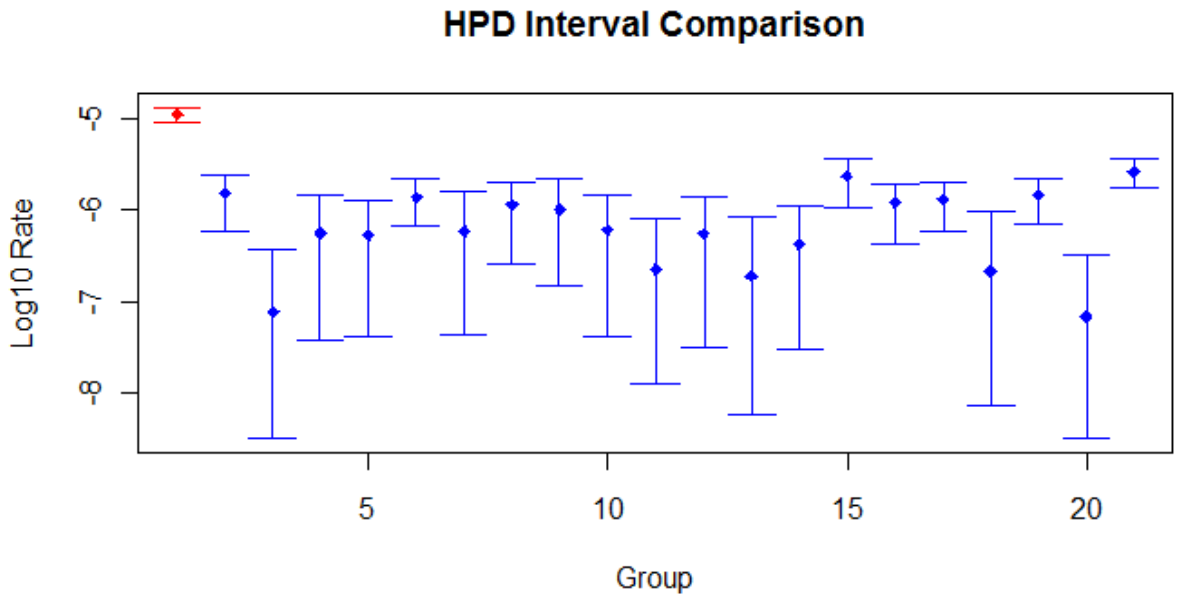

Table S1: Isolates used

We studied 12 field isolates of *Mycoplasma* collected from House Finches in the Southeastern United States. The isolates were chosen to encompass the complete time span of the epizootic with four samples from the 1994-1996 period, four from 2001 and four from 2007. We also studied four isolates of *Mycoplasma* collected from poultry. Table S1 below shows the source of the isolates, their sequencing coverage in terms of the reference genome[1], and any alternate names the strains may have had in previous studies.

Table S1. Characteristics of MG isolates used in study

Strain Name	Host species*	Coverage	Avg. Quality Score	Date Isolated	Isolated From	Source	Alternate Name
AL_2001_13	HF	11.4	27	March 6, 2001	Lee County, Alabama	This study	
AL_2001_17	HF	8.9	18	June 27, 2001	Lee County, Alabama	This study	
AL_2001_53	HF	6.5	24	March 14, 2001	Lee County, Alabama	This study	
AL_2001_61	HF	9.5	23	February 11, 2001	Lee County, Alabama	This study	
AL_2007_05	HF	8.4	34	January 20, 2007	Lee County, Alabama	This study	
AL_2007_10	HF	4.3	37	January 20, 2007	Lee County, Alabama	This study	
AL_2007_37	HF	18.9	23	February 11, 2007	Lee County, Alabama	This study	
AL_2007_38	HF	9.8	35	February 11, 2007	Lee County, Alabama	This study	
GA_1995	HF	7.2	27	February 13, 1995	Clarke County, Georgia	[17]	K3891
KY_1996	HF	7.3	22	February 26, 1996	Kentucky	[18]	K4117
TN_1996	HF	6.8	24	January 23, 1996	Shelby County, TN	[18]	K4094
VA_1994	HF	13.9	24	June, 1994	Virginia	[19]	S11
TK_2001	Turkey	294	33.4	2001	Indiana	[5]	K5054TK01
TK_1998	Turkey	391	33.4	1998	Colorado	[5]	K4669ATK98
TK_1996	Turkey	498	33.4	1996	Missouri	[5]	K4158CTK96
CK_1996	Chicken	460	33.4	1996	Missouri	[5]	K4280CK96

*HF = House Finch

The House Finch isolates from 2001 and 2007 were obtained for this study as follows. House finches were caught in wire mesh cages placed around feeders and in mist nets. Upon capture, *Mycoplasma* samples were collected by swabbing eye conjunctiva and choanal cleft of birds displaying symptoms of disease. Swabs were immediately placed into 3 mL of SP4 media preheated to 37 C. After gentle vortexing, the swab was removed and the inoculated broth [20] was incubated at 37 C overnight. After approximately 24 hours, a 1:10 blind passage was performed for each culture which was then incubated at 37 C for 5 weeks or until a color change indicated growth [21]. Following a media color change, stocks of each isolate were made as follows: 500uL of a 1:1 solution of SP4 broth and glycerol was added to 500uL of cell culture. Samples were gently mixed and frozen at -80 C for long-term storage. DNA for sequencing was prepared by re-inoculating frozen cultures into SP4 media incubated until log phase. DNA was extracted from each sample at between passage five and seven using Qiagen DNA tissue minipreps.

Table S2. SNP counts in the final working data set comprising the 17-way alignment

	All	House Finch Strains	House Finch Strains and TK_2001	1994-1996 Strains	2001 Strains	2001 Strains excluding AL_2001_17	2007 Strains	New Poultry Strains and Reference Genome excluding TK_2001	New Poultry Strains
Total SNPs	16,398	412	469	136	152	42	37	14,400	13,175
Synonymous	9,383	122	138	37	50	12	11	8,459	7,735
Non-synonymous	5,324	246	279	85	88	24	21	4,534	4,090
Non-coding	1,729	45	53	14	15	7	5	1,441	1,377
Singletons	8,576	258	310	115	103	42	36	8,208	5,601
Phylogenetically informative within the group	7,693	152	157	20	48	0	0	6,048	7,517
Fixed SNPS (Ignoring missing data)	N/A	80	1,579	1	3	29	87	1,551	140
Fixed SNPS (Require data from all group members)	N/A	8	310	0	0	20	47	1,459	55
Fixed SNPS (Require data from all strains in study)	N/A	0	301	0	0	9	24	297	0
Fixed SNPS (Require data from all non-group members, but allows incomplete data within specified group)	N/A	2	1,485	0	0	12	36	306	0

Table S3. Sanger Sequencing Validation of SNP Calls

In the early stages of the project we validated a small subset (n = 9) of SNPs via PCR and direct sequencing of 76 sequencing reactions spread across the 12 House Finch strains. We selected these sequenced positions for two reasons. First, these sites were phylogenetically informative for the pre-2001 House Finch strains whose relationships we wished to resolve. Second, we felt these SNPs were the most suspect of all of the SNPs in our dataset as they provided conflicting phylogenetic information and so were either strong evidence for an unknown source of sequencing errors in our methods or strong evidence for recombination in this population of *Mycoplasma*. We were able to rule out sequencing error as all sequenced loci confirmed the polymorphisms identified by the 454 sequencing (71 of 76 loci matched the 454 sequencing data, and 5 of 76 provided data for strains that did not have adequate coverage at that position in the original 454 data).

Table S3. SNPs Validated by PCR amplification and Sanger Sequencing

Strain	Position	Sanger bp	Reference bp	454 bp
AL_2001_13	170360	C	C	C
AL_2001_17	170360	T	C	T
AL_2001_61	170360	C	C	C
AL_2007_05	170360	T	C	T
AL_2007_10	170360	T	C	T
AL_2007_37	170360	T	C	T
AL_2007_38	170360	T	C	T
GA_1995	170360	C	C	C
KY_1996	170360	C	C	C
TN_1996	170360	C	C	N
VA_1994	170360	C	C	C
AL_2001_13	174643	C	C	C
AL_2001_17	174643	C	C	C
AL_2001_53	174643	C	C	C
AL_2001_61	174643	C	C	C
AL_2007_05	174643	T	C	T
AL_2007_10	174643	T	C	T
AL_2007_37	174643	T	C	T
AL_2007_38	174643	T	C	T
GA_1995	174643	C	C	C
KY_1996	174643	T	C	T
TN_1996	174643	C	C	C
VA_1994	174643	C	C	C
AL_2001_13	580857	G	G	G
AL_2001_61	580857	G	G	G
AL_2007_05	580857	T	G	T
AL_2007_37	580857	T	G	T
AL_2001_13	691180	G	G	G

AL_2001_17	691180	A	G	A
AL_2001_61	691180	G	G	G
AL_2001_61	691180	G	G	G
AL_2007_05	691180	A	G	N
AL_2007_10	691180	A	G	A
AL_2007_37	691180	A	G	A
AL_2007_38	691180	A	G	A
GA_1995	691180	G	G	G
KY_1996	691180	G	G	G
TN_1996	691180	G	G	N
AL_2001_17	716811	C	C	C
AL_2007_05	716811	C	C	C
AL_2007_37	716811	C	C	C
AL_2007_38	716811	C	C	C
TN_1996	716811	C	C	C
VA_1994	716811	C	C	C
AL_2001_13	720901	T	T	T
AL_2001_17	720901	T	T	T
AL_2001_53	720901	T	T	T
AL_2001_61	720901	T	T	T
GA_1995	720901	T	T	T
TN_1996	720901	T	T	T
AL_2001_13	853947	A	G	A
AL_2001_17	853947	G	G	G
AL_2001_53	853947	A	G	A
AL_2001_61	853947	A	G	A
AL_2007_05	853947	G	G	G
AL_2007_10	853947	G	G	N
AL_2007_37	853947	G	G	G
AL_2007_38	853947	G	G	G
GA_1995	853947	G	G	G
KY_1996	853947	A	G	A
TN_1996	853947	G	G	G
VA_1994	853947	A	G	A
AL_2001_13	909457	A	C	A
AL_2001_61	909457	A	C	A
AL_2007_05	909457	C	C	C
AL_2007_37	909457	C	C	C
AL_2007_38	909457	C	C	C
VA_1994	909457	C	C	C
AL_2001_13	973203	G	G	G
AL_2001_17	973203	G	G	G
AL_2001_53	973203	G	G	N
AL_2001_61	973203	G	G	G
AL_2007_37	973203	A	G	A
AL_2007_38	973203	A	G	A
GA_1995	973203	G	G	G
VA_1994	973203	G	G	G

.

Table S4. Cross validation of the 454 SNP calls using the Illumina SNP calls

Alignment File Name **Stringent_2010_Masked_4_Val.fna**

Strain	TN_1996	GA_1995	KY_1996	VA_1994	AL_2001_53	AL_2001_17	AL_2001_61	AL_2001_13	AL_2007_10	AL_2007_05	AL_2007_38	AL_2007_37
Differences from Reference	3129	2613	5206	6482	4044	5682	5327	5770	3772	3260	5766	6732
Differences Shared with TK_2001	3121	2604	5194	6460	4022	5621	5292	5737	3723	3212	5694	6642
% Identical SNP calls	99.7%	99.7%	99.8%	99.7%	99.5%	98.9%	99.3%	99.4%	98.7%	98.5%	98.8%	98.7%
Singletons for Strain	6	7	6	15	3	54	9	10	3	3	2	7

Alignment File Name **Stringent_Moderate_v2_2010_Masked_4_Val.fna**

Strain	TN_1996	GA_1995	KY_1996	VA_1994	AL_2001_53	AL_2001_17	AL_2001_61	AL_2001_13	AL_2007_10	AL_2007_05	AL_2007_38	AL_2007_37
Differences from Reference	5402	4763	7012	7482	6118	7269	7120	7347	5722	5379	7181	7428
Differences Shared with TK_2001	5352	4690	6972	7437	6045	7173	7053	7282	5628	5275	7067	7307
% Identical SNP calls	99.1%	98.5%	99.4%	99.4%	98.8%	98.7%	99.1%	99.1%	98.4%	98.1%	98.4%	98.4%
Singletons for Strain	26	58	15	17	29	65	9	5	9	21	7	4

Alignment File Name **Moderate_2010_Masked_4_Val.fna**

Strain	TN_1996	GA_1995	KY_1996	VA_1994	AL_2001_53	AL_2001_17	AL_2001_61	AL_2001_13	AL_2007_10	AL_2007_05	AL_2007_38	AL_2007_37
Differences from Reference	6411	5875	7336	7682	6719	7526	7487	7638	6191	6380	7439	7598
Differences Shared with TK_2001	6306	5699	7262	7615	6553	7399	7374	7545	6017	6186	7291	7456
% Identical SNP calls	98.4%	97.0%	99.0%	99.1%	97.5%	98.3%	98.5%	98.8%	97.2%	97.0%	98.0%	98.1%
Singletons for Strain	70	151	36	25	108	81	38	16	77	78	16	6

Table S5. Estimates of genetic diversity (π) in subgroups of MG strains sampled from different host species* in the LS-MSA

Host species, year	N	bp	π	Standard Deviation
All	73	~1362	0.01963	0.00106
Chicken, all	26	~1362	0.01888	0.00171
Chicken, 1994-1996, inclusive	4	~1362	0.01853	0.00397
Chicken, 1994-1996 (no Australia samples)	2	~1362	0.02428	0.01214
Chicken, post-1996	18	~1362	0.01737	0.00191
All turkey	31	~1362	0.02253	0.00193
Turkey, all	33	~1362	0.02203	0.00159
Turkey,1994-1996, inclusive	10	~1362	0.01634	0.00161
Turkey, post-1996	21	~1362	0.02332	0.00201
House finch, all	14	~1362	0.00057	0.00019
House Finch, this study	**12**	**743,011**	**0.00014**	**0.00001**
1994-1996	**4**	**743,011**	**0.00010**	**0.00003**
2001	**4**	**743,011**	**0.00011**	**0.00004**
2007	**4**	**743,011**	**0.00003**	**0.00001**

Data from this study (bold) and from Ferguson et al. 2005 (5).

Table S6: Patterns of synonymous and nonsynonymous substitutions

We compared the frequencies of non-synonymous, synonymous and non-protein coding SNPs in the House Finch and poultry populations by comparing three groups of SNPs. The first type were polymorphisms that likely arose in the House Finch MG lineage, as they are fixed in the poultry MG strains but are polymorphic amongst the House Finch ones. The second group are those SNPs that likely arose in the poultry MG population, as they show the opposite pattern and are fixed in the House Finch strains. We also examined SNPs that represented fixed differences between the two populations and likely arose on the lineage separating the poultry and House Finch MG populations. 35 SNPs were excluded from categorization because they were polymorphic in both the poultry and House Finch populations. Finally, we obtained expected numbers of the three types of mutations by simulating mutations in the genome using the maximum a posteriori parameters for the HKY substitution model inferred from our earlier BEAST analysis (Text S2).

Observed and expected number of SNPs in various comparisons among strains.

	Polymorphic within House Finch strains	Polymorphic within poultry strains	Fixed Differences	Simulated
Synonymous	28.5%	58.0%	27.8%	25.7%
Nonsynonymous	59.9%	31.6%	40.5%	63.8%
Non-protein Coding	11.6%	10.4%	31.6%	10.5%
Total SNPs	379	15,940	79	10,000

By converting table S6 into a contingency table, one can reject the assumption that the mutations are distributed as one would expect under neutrality as defined by the simulated distribution in the poultry population, but not in the House Finch population. ($p_{poultry}$<2.2e-16, p_{HF}=.28), which is consistent with other studies that have shown very recently diverged pathogens tend to evolve neutrally [22].

To obtain estimates of the distribution of dN/dS values for each gene within MG from all of our samples using PAML v. 4.2b[23]. For each gene, we used the maximum clade credibility tree from our BEAST analysis (Test S2) and for those genes that contained both non-synonymous and synonymous mutations we used PAML to estimate the dn/ds (omega) ratio. These data are summarized in Fig. 2c of the main text.

Table S7. Regions of the reference genome that had been lost in House Finch MG isolates

We searched for genes in the reference genome that were not present in the House Finch MG isolates. The 454 contigs assembled from our pooled 2007 samples were mapped to the reference genome using Megablast and any portion that aligned with greater than 95% similarity and over 100 bp in length to a section of the reference genome was considered to represent that section. We then searched for any section of the reference genome longer than 200 bp in length that was not represented by some of the reads in our sample. Unrepresented segments were then further investigated to confirm the deletion, determine the likely mechanism by which it occurred and the starting and ending points in the coordinates given by the reference genome. For the reasons given previously, any putative deletions that appeared in the VlhA regions were not investigated further in this study, even though these regions likely experienced deletions relative to the reference MG strain.

The list of reconstructed deletions in House Finch MG isolates from this analysis is shown in the Table S7; in total they account for ~42 kb of the reference genome being lost and are responsible for the deletion of a total of 34 genes. Three of these deletions are hypothesized to have occurred via recombination between IS elements. Two of the large deletions (numbers 3 and 5) could clearly be identified because no reads representing the deleted sequence were present and because a contig could be formed that spanned the deletion. However, three of the deletions (numbers 1, 3 and 5) were clearly mediated by an IS element insertion followed by a non-homologous recombination-mediated deletion. As these events are caused by recombination between non-homologous sequences, the exact location of the recombination point is unknown and only approximate coordinates are given in the Table S7. All of the deletions found were present in all of our other HF strains, except for the 12.7 kb deletion which was unique to the 2007 isolates.

Deletion number	Approx. start	Approx. end	Deletion size (bp)	Deletion mediated by recombination between IS elements?	Distribution
1	124,815	126,674	1,859	Yes	All strains except R_{low}
2	137,173	138,833	1,660	No	All strains except R_{low} and CK_1996
3	369,420	388,013	18,593	Yes	All strains except R_{low} and TK_1996
4	912,433	925,150	12,717	No	Only in 2007 House Finch strains
5	938,560	945,976	7,416	Yes	All strains except R_{low}
Total deleted: ~42,245 bp					

Table S8. Descriptions of six novel insertion sites of IS elements and insert characteristics for House Finch MG strains.

	Approximate Location	Sides Present	Target Gene	Description of Insertion Area
A	124818	5’	MGA_0801	Potential C-terminal fragment of subtilisin like protease
B	295023	Both	None	This section of the genome is unannotated. The location is 1,047 and 276 bp away from the genes on either side.
C	464795	Both	MGA_1220	ArcA, a predicted arginine deiminase
D	537089	Both	None	This landed inside a pseudo-gene that formerly was an acetyl-CoA hydrolase/transferase
E	560163	Both	None	This is 201 bp and 167 bp away from the nearest genes on either side.
F	938560	5’	None	This is 142 bp and 151 bp away from the genes on either side of this insertion.

Table S9. Genes pseudogenized or deleted in the House Finch MG isolates and their status in other *Mycoplasma* genomes.

Among the 12 House Finch isolates we identified 34 genes that had been removed by a deletion, 2 that had been disrupted by a transposon insertion (including one that was deleted following this insertion) and 17 genes that had been pseudogenized by frameshift or nonsense mutations, for a total of 52 genes . We sought to evaluate if these genes were unique to the reference MG genome by evaluating if they had any homologues in any of the 20 Mollicute genomes available has determined by the Molligen Database [24]. We found that 5 of the 33 genes (15%) lost by a deletion lacked a homologue in at least one other genome, while 3 of the 17 genes lost by pseudogenization in the House Finch isolates (~18%) lacked a homologue in the other genomes. We also checked whether any of the genes that were lost in the House Finch isolates had homologues in every one of the 13 *Mycoplasma* genomes available in the database, and thus could be considered "core" genes. We found that of the 229 genes in the reference genome that had a homologue in all of the other genomes, 7 of these had been lost by a combination of 1 deletion and 3 frameshift mutations in the House Finch MG strains.

Table S9. Genes pseudogenized or deleted in the House Finch MG isolates and their status in other *Mycoplasma* genomes.

Gene ID	Start	End	How Lost	No Homology to other *Mycoplasma* genomes	Homology to all other *Mycoplasma* genomes	Gene Name	Product
MGA_0625a	5159	6077	Disruptor Mutation	FALSE	FALSE		ABC-type multidrug/protein/lipid (MdlB-like) transport system component domain protein
MGA_0626	6392	8294	Disruptor Mutation	FALSE	TRUE		ABC-type multidrug/protein/lipid (MdlB-like) transport system component
MGA_0641	14480	15212	Disruptor Mutation	FALSE	FALSE	glpF	glycerol uptake facilitator protein GlpF
MGA_0648	18209	20462	Disruptor Mutation	FALSE	FALSE		conserved lipoprotein
MGA_0656	30264	30948	Disruptor Mutation	TRUE	FALSE		unique hypothetical lipoprotein
MGA_0686	50592	52587	Disruptor Mutation	FALSE	TRUE	uvrB	excinuclease ABC subunit B
MGA_0687	52631	53789	Disruptor Mutation	FALSE	FALSE	pstS	ABC-type phosphate transport system periplasmic phosphate binding protein
MGA_0801	124459	125605	Deletion/IS Insertion	FALSE	FALSE		Subtilisin-like serine protease domain protein
MGA_0802	125682	126432	Deletion	TRUE	FALSE		Subtilisin-like serine protease domain protein
MGA_0815	137104	139078	Deletion	FALSE	FALSE		Subtilisin-like serine protease
MGA_1037	332335	334084	Disruptor Mutation	FALSE	FALSE		conserved hypothetical membrane protein
MGA_1328	369554	369794	Deletion	FALSE	TRUE	deoC_1	Deoxyribose-phosphate aldolase domain protein
MGA_1081	369839	371102	Deletion	FALSE	FALSE		putative transposase

MGA_1083	371070	371919	Deletion	FALSE	FALSE		HAD superfamily hydrolase Cof
MGA_1085	371929	373567	Deletion	FALSE	FALSE		conserved hypothetical protein
MGA_1087	373576	374212	Deletion	FALSE	FALSE		conserved hypothetical protein
MGA_1088	374241	374925	Deletion	FALSE	TRUE		ABC transporter ATPase component
MGA_1089	374908	376459	Deletion	FALSE	FALSE		ABC transporter permease domain protein
MGA_1091	376555	376891	Deletion	FALSE	FALSE		putative signal peptidase I
MGA_1092	376969	377530	Deletion	FALSE	TRUE		Elongation factor G domain protein
MGA_1100	379435	380476	Deletion	FALSE	TRUE	asnS_2	Asparaginyl-tRNA synthetase
MGA_1102	380479	382111	Deletion	FALSE	FALSE		conserved hypothetical membrane protein
MGA_1103	382094	384026	Deletion	FALSE	FALSE		predicted integral membrane methylase-domain protein
MGA_1347	384502	384670	Deletion	FALSE	FALSE		putative transposase domain protein
MGA_1106	384754	384946	Deletion	TRUE	FALSE		putative transposase domain protein
MGA_1107	384995	386495	Deletion	FALSE	FALSE		conserved hypothetical RmuC-domain protein
MGA_1108	386618	387119	Deletion	FALSE	FALSE		putative transposase domain protein
MGA_1109	387260	388307	Deletion	FALSE	FALSE		putative transposase domain protein
MGA_1220	464277	465489	IS Insertion	FALSE	FALSE	arcA_1	Arginine deiminase
MGA_1263	507145	507823	Disruptor Mutation	FALSE	FALSE	beta-pgm	putative beta-phosphoglucomutase (beta-PGM)
MGA_1283	520043	520808	Disruptor Mutation	FALSE	FALSE		PTS system mannitol-specific (MtlA)-like IIB domain protein
MGA_1305	536425	536824	Disruptor Mutation	FALSE	FALSE	maoC	MaoC-like dehydratase
MGA_0135	652030	653536	Disruptor Mutation	FALSE	TRUE	potA	ABC-type spermidine/putrescine import ATP-binding protein potA
MGA_0137	653891	655376	Disruptor Mutation	TRUE	FALSE		unique hypothetical protein
MGA_1361	747656	747986	Disruptor Mutation	FALSE	FALSE		unique hypothetical protein
MGA_1354	876961	877114	Disruptor Mutation	FALSE	FALSE		hypothetical protein
MGA_0508	910763	912797	Deletion	FALSE	FALSE	fruA	PTS system fructose-specific enzyme EIIABC component

MGA_0512	912799	913246	Deletion	TRUE	FALSE		hypothetical protein
MGA_0514	913193	914144	Deletion	FALSE	FALSE	manA	mannose-6-phosphate isomerase (phosphomannose isomerase)
MGA_0516	915226	916669	Deletion	TRUE	FALSE		unique hypothetical protein
MGA_0517	916577	917954	Deletion	FALSE	FALSE		Subtilisin-like serine protease domain protein
MGA_0518	917874	918705	Deletion	TRUE	FALSE		Subtilisin-like serine protease domain protein
MGA_0519	919247	923060	Deletion	FALSE	FALSE		Csn1 family CRISPR-associated protein
MGA_0523	923127	924054	Deletion	FALSE	FALSE	cas1	CRISPR-associated protein Cas1
MGA_0525	924040	924370	Deletion	FALSE	FALSE	cas2	CRISPR-associated protein Cas2
MGA_0526	924369	925134	Deletion	FALSE	FALSE		conserved hypothetical protein
MGA_0537	938710	941338	Deletion	FALSE	FALSE	hsdM	type I restriction-modification system methyltransferase (M) subunit
MGA_0539	941547	942165	Deletion	FALSE	FALSE	hsdS_1	type I restriction-modification system specificity (S) subunit domain protein
MGA_0540	942139	942724	Deletion	FALSE	FALSE	hsdS_2	type I restriction-modification system specificity (S) subunit domain protein
MGA_0541	942734	945890	Deletion	FALSE	FALSE	hsdR	type I site-specific restriction-modification system restriction (R) subunit (deoxyribonuclease)
MGA_0567	970243	970549	Disruptor Mutation	TRUE	FALSE		unique hypothetical protein
MGA_0586	987980	990098	Disruptor Mutation	FALSE	FALSE		conserved hypothetical protein

Table S10: Mutations in the *UvrB* Gene and Possible Effects

The *UvrB* gene in every house finch MG strain sampled contains a mutation that truncates the final 3 amino acids of the protein, and this mutation is also present in the closely related TK_2001. The DNA encoding the C-terminal of this amino acid contains a 2 time repeat of the sequence "TAAG" and this mutation introduced one additional repeat of this sequence as a 4 bp insertion. The effect of this 4 bp insertion was to introduce an early "TAA" stop codon and thereby truncate the protein by 3 amino acids as shown below.

Comparison of the C-Terminals in the *UvrB* gene

House Finch MG Isolates	…KMIEDLRNEMLEAAKNQNYEHAASLRDLIIELETQQLSK*
Reference MG Genome	…KMIEDLRNEMLEAAKNQNYEHAASLRDLIIELETQQLSKTNK*

UvrB is an integral part of the cell's DNA excision repair system and functions by forming associations with *UvrA* and *UvrC* during the repair process. Experimental work with the *UvrB* protein from E. coli has shown that the C-terminal of this protein is essential for the protein to associate with *UvrC* and allow a repair to occur [25,26]. However, the house finch MG protein has lost only the final 3 amino acids, and so the specific effect of this mutation cannot be determined from past functional or comparative work.

DNA excision repair is responsible for the repair of pyrimidine dimers, and one signature that these types of mutations have not been repaired along an evolving lineage is the presence of "CC" to "TT" mutations (or "GG" to "AA" if the effect of the mutation is viewed from the other strand). To investigate if the rate of these mutations is elevated in the house finch MG samples, we compared the characteristics of adjacent SNPs that are found segregating amongst the house finch and TK_2001 MG samples to those adjacent SNPs that are polymorphic amongst the reference genome and the other poultry strains. This comparison is shown in table S14.

This comparison showed many features that suggested inhibition of the nucleotide excision repair system within the house MG. The majority of the double mutations within the house finch MG could be identified as involving a "CC" to "TT" substitution on one of the strands of DNA. Among the house finch MG samples, 14 pairs of SNPs were adjacent to each other (Table S11). Of these, 13 could be parsimoniously identified as having occurred on a single, and the same, branch of the tree, and 12 of these could be defined (using the reference and poultry strains to identify the derived allele) as a "CC" to "TT" substitution. Of the two remaining adjacent SNP pairs (at reference positions 667,905 and 715,595), one involved two mutations that occurred on separate branches on the tree, such that no genotype contained a copy of both derived alleles, and another involved an "AA" to "TT" transition. Also suggestive of an increase in the mutation rate for paired bases is the high number of adjacent SNPs given the small number of total SNPs within the HF samples. The percentage of SNPs that are adjacent to each other is expected to increase with the total number of SNPs in an alignment. However, despite having a much smaller number of SNPs, those that were polymorphic among the house finch MG strains contained a greater proportion of adjacent SNPs (Table S11).

We tested for an increase in the number of paired substitutions that involved a substitution from two identical bases to two identical bases of a different type. A contingency table for this analysis was constructed by counting only the adjacent SNPs that appeared in pairs (excluding SNPs that appeared in adjacent groups of three or more, as well as SNPs with over 2 types segregating). The frequency of identical conversions in each group was then compared and found to be significantly different ($p < 0.00001$). This analysis is slightly complicated because at one of the positions containing adjacent SNPs in the house finch MG samples, position 667,905 in the reference genome coordinates, the ancestral sequence "CC" sequence has mutated in one strain to create a "CC"-> "CA" substitution, while on the branch leading to the 2007 strains it has mutated to create a "CC"-> "TT" substitution. Although this made the classification of this pair ambiguous, the frequency difference for these types of mutations suggests that either classification still results in a significant difference, though for clarity we presented it as an identical pair substitution in table S14.

Table S10 – Comparison of adjacent SNPs within the house finch MG to those between the house finch MG and the reference genome.

	SNPs polymorphic amongst strains without the *UvrB* mutation but fixed amongst strains that have it	**SNPs that are polymorphic amongst the strains with the *UvrB* mutation.**
Total SNPs	16,959	420*
SNPs adjacent to another SNP (percentage of total SNPs)	1,458 (8.5%)	28 (6.8%)
Adjacent Pairs of SNPs (excluding >3 SNPs in a row)	641	14
Adjacent pairs with a conversion of an identical pair to an identical pair (e.g. "CC"->"TT"); (percentage of total adjacent pairs)	42 (6.6%)	13 (92.8%)
Adjacent pairs with non-identical conversions (eg. "AA"->"TC", "AT"->"GC" or "GC" ->"CC") (percentage of total adjacent pairs)	599 (93.4%)	1 (7.2%)

Table S11. Instances of polymorphic adjacent SNPs among the house finch MG strains.

	Position of double SNP in Reference Coordinates													
Strain	14,966	61,514	76,728	120,043	169,641	225,915	241,224	303,492	315,466	572,038	667,905	688,985	715,595	803,438
R Low	GG	CC	GG	GG	CC	CC	AA	GG	CC	GG	CC	CC	TC	GG
TN_1996		CC	GG	GG	CC	CC	AA	GG	CC	GG	CC	CC	TC	GG
VA_1994	GG	CC	GG	GG	CC	CC	AA	GG	CC	GG	CC	CC	TC	AA
KY_1996	GG	CC	GG	GG	CC	CC	AA	GG	CC	GG	CC	CC	TC	AA
GA_1995	GG	CC	GG	GG	CC	CC	AA	GG	TT	GG	CC	CC	TC	AA
AL_2001_53	GG	CC	GG	GG	CC	CC	AA	GG	CC	GG	CA	CC	TT	AA
AL_2001_17	AA	CC	AA	GG	CC	CC	AA	AA	CC	AA	TT	TT	CC	AA
AL_2001_61	GG	CC	GG	AA	CC	CC	AA	GG	CC	GG	CA	CC	TT	AA
AL_2001_13	GG	CC	GG	GG	CC	CC	AA	GG	CC	GG	CA	CC	TT	AA
AL_2007_10	GG	TT	GG	GG	TT	TT	TT	GG	CC	GG	CA	CC	TT	AA
AL_2007_05	GG	TT	GG	GG	TT	TT	TN	GG	CC	GG	CA	CC	TT	AA
AL_2007_38	GG	TT	GG	GG	TT	TT	TN	GG	CC	GG	CA	CC	TT	AA
AL_2007_37	GG	TT	GG	GG	TT	TT	TN	GG	CC	GG	CA	CC	TT	AA
TK_2001	GG	CC	GG	GG	CC	CC	AA	GG	CC	GG	CC	CC	TT	AA
CK_1996	GG	CC	GG	GG	CC	CC	AA	GG	CC	GG	CC	CC	TC	GG
TK_1998	GG	CC	GG	GG	CC	CC	AA	GG	CC	GG	CC	CC	TC	GG
TK_1996	GG	CC	GG	GG	CC	CC	AA	GG	CC	GG	CC	CC	TC	GG

Table S12. Counts of unique and total (due to duplication) CRISPR spacers from each strain.

Strain	Total Unique Spacers	Total Spacers
CK_1996	66	75
TK_1998	35	36
TK_1996	93	147
Reference Genome	61	71
TK_2001	38	42
VA_1994	34	36
TN_1995	38	40
KY_1995	35	39
GA_1995	47	50
AL_2001_17	37	37
AL_2001_53	40	40
AL_2001_61	40	40
AL_2001_13	39	39
AL_2007_10	28	28
AL_2007_05	28	28
AL_2007_38	28	29
AL_2007_37	28	28
All Strains	**302**	**805**